

Electronic Journal of Statistics

Vol. 6 (2012) 354–381

ISSN: 1935-7524

DOI: [10.1214/12-EJS675](https://doi.org/10.1214/12-EJS675)

The Bernstein-Von-Mises theorem under misspecification

B.J.K. Kleijn* and A.W. van der Vaart

*Korteweg-de Vries Institute for Mathematics, Universiteit van Amsterdam
Mathematical Institute, Universiteit Leiden*

e-mail: B.Kleijn@uva.nl; avdvaart@math.leidenuniv.nl

Abstract: We prove that the posterior distribution of a parameter in misspecified LAN parametric models can be approximated by a random normal distribution. We derive from this that Bayesian credible sets are not valid confidence sets if the model is misspecified. We obtain the result under conditions that are comparable to those in the well-specified situation: uniform testability against fixed alternatives and sufficient prior mass in neighbourhoods of the point of convergence. The rate of convergence is considered in detail, with special attention for the existence and construction of suitable test sequences. We also give a lemma to exclude testable model subsets which implies a misspecified version of Schwartz’ consistency theorem, establishing weak convergence of the posterior to a measure degenerate at the point at minimal Kullback-Leibler divergence with respect to the true distribution.

AMS 2000 subject classifications: Primary 62F15, 62F25, 62F12.

Keywords and phrases: Misspecification, posterior distribution, credible set, limit distribution, rate of convergence, consistency.

Received December 2010.

Contents

1	Introduction	355
2	Posterior limit distribution	357
2.1	Asymptotic normality in LAN models	357
2.2	Asymptotic normality in the i.i.d. case	360
2.3	Asymptotic normality of point-estimators	362
3	Rate of convergence	365
3.1	Posterior rate of convergence	366
3.2	Suitable test sequences	370
4	Consistency and testability	376
4.1	Exclusion of testable model subsets	376
5	Appendix: Technical lemmas	379
	References	380

*Research supported by a VENI-grant, Netherlands Organisation for Scientific Research.

1. Introduction

The Bernstein-Von Mises theorem asserts that the posterior distribution of a parameter in a smooth finite-dimensional model is approximately a normal distribution if the number of observations tends to infinity. Apart from having considerable philosophical interest, this theorem is the basis for the justification of Bayesian credible sets as valid confidence sets in the frequentist sense: (central) sets of posterior probability $1 - \alpha$ cover the parameter at confidence level $1 - \alpha$. In this paper we study the posterior distribution in the situation that the observations are sampled from a “true distribution” that does not belong to the statistical model, i.e. the model is misspecified. Although consistency of the posterior distribution and the asymptotic normality of the Bayes estimator (the posterior mean) in this case have been considered in the literature, by Berk (1966, 1970) [2, 3] and Bunke and Milhaud (1998) [5], the behaviour of the full posterior distribution appears to have been neglected. This is surprising, because in practice the assumption of correct specification of a model may be hard to justify. In this paper we derive the asymptotic normality of the full posterior distribution in the misspecified situation under conditions comparable to those obtained by Le Cam in the well-specified case.

This misspecified version of the Bernstein-Von Mises theorem has an important consequence for the interpretation of Bayesian credible sets. In the misspecified situation the posterior distribution of a parameter shrinks to the point within the model at minimum Kullback-Leibler divergence to the true distribution, a consistency property that it shares with the maximum likelihood estimator. Consequently one can consider both the Bayesian procedure and the maximum likelihood estimator as estimates of this minimum Kullback-Leibler point. A confidence region for this minimum Kullback-Leibler point can be built around the maximum likelihood estimator based on its asymptotic normal distribution, involving the sandwich covariance. One might also hope that a Bayesian credible set automatically yields a valid confidence set for the minimum Kullback-Leibler point. However, the misspecified Bernstein-Von Mises theorem shows the latter to be false.

More precisely, let $B \mapsto \Pi_n(B \mid X_1, \dots, X_n)$ be the posterior distribution of a parameter θ based on observations X_1, \dots, X_n sampled from a density p_θ and a prior measure Π on the parameter set $\Theta \subset \mathbb{R}^d$. The Bernstein-Von Mises theorem asserts that if X_1, \dots, X_n is a random sample from the density p_{θ_0} , the model $\theta \mapsto p_\theta$ is appropriately smooth and identifiable, and the prior puts positive mass around the parameter θ_0 , then,

$$P_{\theta_0}^n \sup_B \left| \Pi_n(B \mid X_1, \dots, X_n) - N_{\hat{\theta}_n, (ni_{\theta_0})^{-1}}(B) \right| \rightarrow 0,$$

where $N_{x, \Sigma}$ denotes the (multivariate) normal distribution centred on x with covariance matrix Σ , $\hat{\theta}_n$ may be any efficient estimator sequence of the parameter and i_θ is the Fisher information matrix of the model at θ . It is customary to identify $\hat{\theta}_n$ as the maximum likelihood estimator in this context (correct under regularity conditions).

The Bernstein-Von Mises theorem implies that any random sets \hat{B}_n such that $\Pi_n(\hat{B}_n \mid X_1, \dots, X_n) = 1 - \alpha$ for each n satisfy,

$$N_{0,I}((ni_{\theta_0})^{1/2}(\hat{B}_n - \hat{\theta}_n)) \rightarrow 1 - \alpha,$$

in probability. In other words, such sets \hat{B}_n can be written in the form $\hat{B}_n = \hat{\theta}_n + i_{\theta_0}^{-1/2} \hat{C}_n / \sqrt{n}$ for sets \hat{C}_n that receive asymptotically probability $1 - \alpha$ under the standard Gaussian distribution. This shows that the $1 - \alpha$ -credible sets \hat{B}_n are asymptotically equivalent to the Wald $1 - \alpha$ -confidence sets based on the asymptotically normal estimators $\hat{\theta}_n$, and consequently they are valid $1 - \alpha$ confidence sets.

In this paper we consider the situation that the posterior distribution is formed in the same way relative to a model $\theta \mapsto p_\theta$, but we assume that the observations are sampled from a density p_0 that is not necessarily of the form p_{θ_0} for some θ_0 . We shall show that the Bernstein-Von Mises can be extended to this situation, in the form,

$$P_0^n \sup_B \left| \Pi_n(B \mid X_1, \dots, X_n) - N_{\hat{\theta}_n, (nV_{\theta^*})^{-1}}(B) \right| \rightarrow 0,$$

where θ^* is the parameter value θ minimizing the Kullback-Leibler divergence $\theta \mapsto P_0 \log(p_0/p_\theta)$ (provided it exists and is unique, see corollary 4.2), V_{θ^*} is minus the second derivative matrix of this map, and $\hat{\theta}_n$ are suitable estimators.

Under regularity conditions the estimators $\hat{\theta}_n$ can again be taken equal to the maximum likelihood estimators (for the misspecified model), which typically satisfy that the sequence $\sqrt{n}(\hat{\theta}_n - \theta^*)$ is asymptotically normal with mean zero and covariance matrix given by the “sandwich formula” $\Sigma_{\theta^*} = V_{\theta^*} (P_0 \dot{\ell}_{\theta^*} \dot{\ell}_{\theta^*}^T)^{-1} V_{\theta^*}$. (See, for instance, Huber (1967) [8] or Van der Vaart (1998) [18].) The usual confidence sets for the misspecified parameter take the form $\hat{\theta}_n + \Sigma_{\theta^*}^{1/2} C / \sqrt{n}$ for C a central set in the Gaussian distribution. Because the covariance matrix V_{θ^*} appearing in the misspecified Bernstein-Von Mises theorem is not the sandwich covariance matrix, central posterior sets of probability $1 - \alpha$ do not correspond to these misspecified Wald sets. Although they are correctly centered, they may have the wrong width, and are in general not $1 - \alpha$ -confidence sets. We show below by example that the credible sets may over- or under-cover, depending on the true distribution of the observations and the model, and to extreme amounts.

The first results concerning limiting normality of a posterior distribution date back to Laplace (1820) [10]. Later, Bernstein (1917) [1] and Von Mises (1931) [14] proved results to a similar extent. Le Cam used the term ‘Bernstein-Von Mises theorem’ in 1953 [11] and proved its assertion in greater generality. Walker (1969) [19] and Dawid (1970) [6] gave extensions to these results and Bickel and Yahav (1969) [4] proved a limit theorem for posterior means. A version of the theorem involving only first derivatives of the log-likelihood in combination with testability and prior mass conditions (compare with Schwartz’ consistency theorem, Schwartz (1965) [16]) can be found in Van der Vaart (1998) [18] which copies (and streamlines) the approach Le Cam presented in [12].

Weak convergence of the posterior distribution to the degenerate distribution at θ^* under misspecification was shown by Berk (1966, 1970) [2, 3], while Bunke and Milhaud (1998) [5] proved asymptotic normality of the posterior mean. These authors also discuss the situation that the point of minimum Kullback-Leibler divergence may be non-unique, which obviously complicates the asymptotic behaviour considerably. Posterior rates of convergence in misspecified non-parametric models were considered in Kleijn and Van der Vaart (2006) [9].

In the present paper we address convergence of the full posterior under mild conditions comparable to those in Van der Vaart (1998) [18]. The presentation is split into two parts. In section 2 we derive normality of the posterior given that it shrinks at a \sqrt{n} -rate of posterior convergence (theorem 2.1). We actually state this result for the general situation of locally asymptotically normal (LAN) models, and next specify to the *i.i.d.* case. Next in section 3 we discuss results guaranteeing the desired rate of convergence, where we first show sufficiency of existence of certain tests (theorem 3.1), and next construct appropriate tests (theorem 3.3). We conclude with a lemma (applicable in parametric and non-parametric situations alike) to exclude testable model subsets, which implies a misspecified version of Schwartz' consistency theorem.

In subsection 2.1 we work in a general locally asymptotically normal set-up, but in the remainder of the paper we consider the situation of *i.i.d.* observations, considered previously and described precisely in section 2.2.

2. Posterior limit distribution

2.1. Asymptotic normality in LAN models

Let Θ be an open subset of \mathbb{R}^d parameterising statistical models $\{P_\theta^{(n)} : \theta \in \Theta\}$. For simplicity, we assume that for each n there exists a single measure that dominates all measures $P_\theta^{(n)}$ as well as a “true measure” $P_0^{(n)}$, and we assume that there exist densities $p_\theta^{(n)}$ and $p_0^{(n)}$ such that the maps $(\theta, x) \mapsto p_\theta^{(n)}$ are measurable.

We consider models satisfying a stochastic local asymptotic normality (LAN) condition around a given inner point $\theta^* \in \Theta$ and relative to a given norming rate $\delta_n \rightarrow 0$: there exist random vectors Δ_{n,θ^*} and nonsingular matrices V_{θ^*} such that the sequence Δ_{n,θ^*} is bounded in probability, and for every compact set $K \subset \mathbb{R}^d$,

$$\sup_{h \in K} \left| \log \frac{p_{\theta^* + \delta_n h}^{(n)}(X^{(n)})}{p_{\theta^*}^{(n)}} - h^T V_{\theta^*} \Delta_{n,\theta^*} - \frac{1}{2} h^T V_{\theta^*} h \right| \rightarrow 0, \quad (2.1)$$

in (outer) $P_0^{(n)}$ -probability. We state simple conditions ensuring this condition for the case of *i.i.d.* observations in section 2.2.

The prior measure Π on Θ is assumed to be a probability measure with Lebesgue-density π , continuous and positive on a neighbourhood of a given

point θ^* . Priors satisfying these criteria assign enough mass to (sufficiently small) balls around θ^* to allow for optimal rates of convergence of the posterior if certain regularity conditions are met (see section 3).

The posterior based on an observation $X^{(n)}$ is denoted $\Pi_n(\cdot | X^{(n)})$: for every Borel set A ,

$$\Pi_n(A | X^{(n)}) = \int_A p_\theta^{(n)}(X^{(n)}) \pi(\theta) d\theta \Big/ \int_\Theta p_\theta^{(n)}(X^{(n)}) \pi(\theta) d\theta. \quad (2.2)$$

To denote the random variable associated with the posterior distribution, we use the notation ϑ . Note that the assertion of theorem 2.1 below involves convergence *in P_0 -probability*, reflecting the sample-dependent nature of the two sequences of measures converging in total-variation norm.

Theorem 2.1. *Assume that (2.1) holds for some $\theta^* \in \Theta$ and let the prior Π be as indicated above. Furthermore, assume that for every sequence of constants $M_n \rightarrow \infty$, we have:*

$$P_0^{(n)} \Pi_n(\|\vartheta - \theta^*\| > \delta_n M_n \mid X^{(n)}) \rightarrow 0. \quad (2.3)$$

Then the sequence of posteriors converges to a sequence of normal distributions in total variation:

$$\sup_B \left| \Pi_n((\vartheta - \theta^*)/\delta_n \in B \mid X^{(n)}) - N_{\Delta_{n,\theta^*}, V_{\theta^*}^{-1}}(B) \right| \xrightarrow{P_0} 0. \quad (2.4)$$

Proof. The proof is split into two parts: in the first part, we prove the assertion conditional on an arbitrary compact set $K \subset \mathbb{R}^d$ and in the second part we use this to prove (2.4). Throughout the proof we denote the posterior for $H = (\vartheta - \theta^*)/\delta_n$ given $X^{(n)}$ by Π_n . The posterior for H follows from that for θ by $\Pi_n(H \in B | X^{(n)}) = \Pi_n((\vartheta - \theta^*)/\delta_n \in B | X^{(n)})$ for all Borel sets B . Furthermore, we denote the normal distribution $N_{\Delta_{n,\theta^*}, V_{\theta^*}^{-1}}$ by Φ_n . For a compact subset $K \subset \mathbb{R}^d$ such that $\Pi_n(H \in K | X^{(n)}) > 0$, we define a conditional version Π_n^K of Π_n by $\Pi_n^K(B | X^{(n)}) = \Pi_n(B \cap K | X^{(n)}) / \Pi_n(K | X^{(n)})$. Similarly we defined a conditional measure Φ_n^K corresponding to Φ_n .

Let $K \subset \mathbb{R}^d$ be a compact subset of \mathbb{R}^d . For every open neighbourhood $U \subset \Theta$ of θ^* , $\theta^* + K\delta_n \subset U$ for large enough n . Since θ^* is an internal point of Θ , for large enough n the random functions $f_n : K \times K \rightarrow \mathbb{R}$,

$$f_n(g, h) = \left(1 - \frac{\phi_n(h)}{\phi_n(g)} \frac{s_n(g)}{s_n(h)} \frac{\pi_n(g)}{\pi_n(h)} \right)_+,$$

are well-defined, with $\phi_n : K \rightarrow \mathbb{R}$ the Lebesgue density of the (randomly located) distribution $N_{\Delta_{n,\theta^*}, V_{\theta^*}}$, $\pi_n : K \rightarrow \mathbb{R}$ the Lebesgue density of the prior for the centred and rescaled parameter H and $s_n : K \rightarrow \mathbb{R}$ the likelihood quotient $s_n(h) = p_{\theta^* + h\delta_n}^{(n)} / p_{\theta^*}^{(n)}(X^{(n)})$.

By the LAN assumption, we have for every random sequence $(h_n) \subset K$, $\log s_n(h_n) = h_n^T V_{\theta^*} \Delta_{n,\theta^*} - \frac{1}{2} h_n^T V_{\theta^*} h_n + o_{P_0}(1)$. For any two sequences $(h_n), (g_n)$

in K , $\pi_n(g_n)/\pi_n(h_n) \rightarrow 1$ as $n \rightarrow \infty$, leading to,

$$\begin{aligned} & \log \frac{\phi_n(h_n)}{\phi_n(g_n)} \frac{s_n(g_n)}{s_n(h_n)} \frac{\pi_n(g_n)}{\pi_n(h_n)} \\ &= (g_n - h_n)^T V_{\theta^*} \Delta_{n,\theta^*} + \frac{1}{2} h_n^T V_{\theta^*} h_n - \frac{1}{2} g_n^T V_{\theta^*} g_n + o_{P_0}(1) \\ & \quad - \frac{1}{2} (h_n - \Delta_{n,\theta^*})^T V_{\theta^*} (h_n - \Delta_{n,\theta^*}) + \frac{1}{2} (g_n - \Delta_{n,\theta^*})^T V_{\theta^*} (g_n - \Delta_{n,\theta^*}) \\ &= o_{P_0}(1), \end{aligned}$$

as $n \rightarrow \infty$. Since all functions f_n depend continuously on (g, h) and $K \times K$ is compact, we conclude that,

$$\sup_{g, h \in K} f_n(g, h) \xrightarrow{P_0} 0, \quad (n \rightarrow \infty), \quad (2.5)$$

where the convergence is in outer probability.

Assume that K contains a neighbourhood of 0 (to guarantee that $\Phi_n(K) > 0$) and let Ξ_n denote the event that $\Pi_n(K) > 0$. Let $\eta > 0$ be given and based on that, define the events:

$$\Omega_n = \left\{ \sup_{g, h \in K} f_n(g, h) \leq \eta \right\}_*,$$

where the $*$ denotes the inner measurable cover set, in case the set on the right is not measurable. Consider the inequality (recall that the total-variation norm $\|\cdot\|$ is bounded by 2):

$$P_0^{(n)} \|\Pi_n^K - \Phi_n^K\| 1_{\Xi_n} \leq P_0^{(n)} \|\Pi_n^K - \Phi_n^K\| 1_{\Omega_n \cap \Xi_n} + 2P_0^{(n)}(\Xi_n \setminus \Omega_n). \quad (2.6)$$

As a result of (2.5) the second term is $o(1)$ as $n \rightarrow \infty$. The first term on the r.h.s. is calculated as follows:

$$\begin{aligned} \frac{1}{2} P_0^{(n)} \|\Pi_n^K - \Phi_n^K\| 1_{\Omega_n \cap \Xi_n} &= P_0^{(n)} \int \left(1 - \frac{d\Phi_n^K}{d\Pi_n^K} \right)_+ d\Pi_n^K 1_{\Omega_n \cap \Xi_n} \\ &= P_0^{(n)} \int_K \left(1 - \int_K \frac{s_n(g)\pi_n(g)\phi_n^K(h)}{s_n(h)\pi_n(h)\phi_n^K(g)} d\Phi_n^K(g) \right)_+ d\Pi_n^K(h) 1_{\Omega_n \cap \Xi_n}. \end{aligned}$$

Note that for all $g, h \in K$, $\phi_n^K(h)/\phi_n^K(g) = \phi_n(h)/\phi_n(g)$, since on K ϕ_n^K differs from ϕ_n only by a normalisation factor. We use Jensen's inequality (with respect to the Φ_n^K -expectation) for the (convex) function $x \mapsto (1-x)_+$ to derive:

$$\begin{aligned} & \frac{1}{2} P_0^{(n)} \|\Pi_n^K - \Phi_n^K\| 1_{\Omega_n \cap \Xi_n} \\ & \leq P_0^{(n)} \int \left(1 - \frac{s_n(g)\pi_n(g)\phi_n(h)}{s_n(h)\pi_n(h)\phi_n(g)} \right)_+ d\Phi_n^K(g) d\Pi_n^K(h) 1_{\Omega_n \cap \Xi_n} \\ & \leq P_0^{(n)} \int \sup_{g, h \in K} f_n(g, h) 1_{\Omega_n \cap \Xi_n} d\Phi_n^K(g) d\Pi_n^K(h) \leq \eta. \end{aligned}$$

Combination with (2.6) shows that for all compact $K \subset \mathbb{R}^d$ containing a neighbourhood of 0, $P_0^{(n)} \|\Pi_n^K - \Phi_n^K\| 1_{\Xi_n} \rightarrow 0$.

Now, let (K_m) be a sequence of balls centred at 0 with radii $M_m \rightarrow \infty$. For each $m \geq 1$, the above display holds, so if we choose a sequence of balls (K_n) that traverses the sequence K_m slowly enough, convergence to zero can still be guaranteed. Moreover, the corresponding events $\Xi_n = \{\Pi_n(K_n) > 0\}$ satisfy $P_0^{(n)}(\Xi_n) \rightarrow 1$ as a result of (2.3). We conclude that there exists a sequence of radii (M_n) such that $M_n \rightarrow \infty$ and $P_0^{(n)} \|\Pi_n^{K_n} - \Phi_n^{K_n}\| \rightarrow 0$ (where it is understood that the conditional probabilities on the *l.h.s.* are well-defined on sets of probability growing to one). The total variation distance between a measure and its conditional version given a set K satisfies $\|\Pi - \Pi^K\| \leq 2\Pi(K^c)$. Combining this with (2.3) and lemma 5.2, we conclude that $P_0^{(n)} \|\Pi_n - \Phi_n\| \rightarrow 0$, which implies (2.4). \square

Condition (2.3) fixes the rate of convergence of the posterior distribution to be that occurring in the LAN property. Sufficient conditions to satisfy (2.3) in the case of *i.i.d.* observations are given in section 3.

2.2. Asymptotic normality in the *i.i.d.* case

Consider the situation that the observation is a vector $X^{(n)} = (X_1, \dots, X_n)$ and the model consists of n -fold product measures $P_\theta^{(n)} = P_\theta^n$, where the components P_θ are given by densities p_θ such that the maps $(\theta, x) \mapsto p_\theta(x)$ are measurable and $\theta \mapsto p_\theta$ is smooth (in the sense of lemma 2.1). Assume that the observations form an *i.i.d.* sample from a distribution P_0 with density p_0 relative to a common dominating measure. Assume that the Kullback-Leibler divergence of the model relative to P_0 is finite and minimized at $\theta^* \in \Theta$, *i.e.*:

$$-P_0 \log \frac{p_{\theta^*}}{p_0} = \inf_{\theta \in \Theta} -P_0 \log \frac{p_\theta}{p_0} < \infty. \quad (2.7)$$

In this situation we set $\delta_n = n^{-1/2}$ and use $\Delta_{n,\theta^*} = V_{\theta^*}^{-1} \mathbb{G}_n \dot{\ell}_{\theta^*}$ as the centering sequence (where $\dot{\ell}_{\theta^*}$ denotes the score function of the model $\theta \mapsto p_\theta$ at θ^* and $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P_0)$ is the empirical process).

Lemmas that establish the LAN expansion (2.1) (for an overview, see, for instance Van der Vaart (1998) [18]) usually assume a well-specified model, whereas current interest requires local asymptotic normality in misspecified situations. To that end we consider the following lemma which gives sufficient conditions.

Lemma 2.1. *If the function $\theta \mapsto \log p_\theta(X_1)$ is differentiable at θ^* in P_0 -probability with derivative $\dot{\ell}_{\theta^*}(X_1)$ and:*

- (i) *there is an open neighbourhood U of θ^* and a square-integrable function m_{θ^*} such that for all $\theta_1, \theta_2 \in U$:*

$$\left| \log \frac{p_{\theta_1}}{p_{\theta_2}} \right| \leq m_{\theta^*} \|\theta_1 - \theta_2\|, \quad (P_0 - a.s.), \quad (2.8)$$

(ii) the Kullback-Leibler divergence with respect to P_0 has a 2nd-order Taylor-expansion around θ^* :

$$-P_0 \log \frac{p_\theta}{p_{\theta^*}} = \frac{1}{2}(\theta - \theta^*)V_{\theta^*}(\theta - \theta^*) + o(\|\theta - \theta^*\|^2), \quad (\theta \rightarrow \theta^*), \quad (2.9)$$

where V_{θ^*} is a positive-definite $d \times d$ -matrix,

then (2.1) holds with $\delta_n = n^{-1/2}$ and $\Delta_{n,\theta^*} = V_{\theta^*}^{-1}\mathbb{G}_n\dot{\ell}_{\theta^*}$. Furthermore, the score function is bounded as follows:

$$\|\dot{\ell}_{\theta^*}(X)\| \leq m_{\theta^*}(X), \quad (P_0 - a.s.). \quad (2.10)$$

Finally, we have:

$$P_0\dot{\ell}_{\theta^*} = \frac{\partial}{\partial\theta}[P_0 \log p_\theta]_{\theta=\theta^*} = 0. \quad (2.11)$$

Proof. Using lemma 19.31 in Van der Vaart (1998) [18] for $\ell_\theta(X) = \log p_\theta(X)$, the conditions of which are satisfied by assumption, we see that for any sequence (h_n) that is bounded in P_0 -probability:

$$\mathbb{G}_n\left(\sqrt{n}(\ell_{\theta^*+(h_n/\sqrt{n})} - \ell_{\theta^*}) - h_n^T\dot{\ell}_{\theta^*}\right) \xrightarrow{P_0} 0. \quad (2.12)$$

Hence, we see that,

$$n\mathbb{P}_n \log \frac{p_{\theta^*+h_n/\sqrt{n}}}{p_{\theta^*}} - \mathbb{G}_n h_n^T \dot{\ell}_{\theta^*} - nP_0 \log \frac{p_{\theta^*+h_n/\sqrt{n}}}{p_{\theta^*}} = o_{P_0}(1).$$

Using the second-order Taylor-expansion (2.9):

$$P_0 \log \frac{p_{\theta^*+h_n/\sqrt{n}}}{p_{\theta^*}} - \frac{1}{2n} h_n^T V_{\theta^*} h_n = o_{P_0}(1),$$

and substituting the log-likelihood product for the first term, we find (2.1). The proof of the remaining assumptions is standard. \square

Regarding the centering sequence Δ_{n,θ^*} and its relation to the maximum-likelihood estimator, we note the following lemma concerning the limit distribution of maximum-likelihood sequences.

Lemma 2.2. Assume that the model satisfies the conditions of lemma 2.1 with non-singular V_{θ^*} . Then a sequence of estimators $\hat{\theta}_n$ such that $\hat{\theta}_n \xrightarrow{P_0} \theta^*$ and,

$$\mathbb{P}_n \log p_{\hat{\theta}_n} \geq \sup_{\theta} \mathbb{P}_n \log p_{\theta} - o_{P_0}(n^{-1}),$$

satisfies the asymptotic expansion:

$$\sqrt{n}(\hat{\theta}_n - \theta^*) = \frac{1}{\sqrt{n}} \sum_{i=1}^n V_{\theta^*}^{-1} \dot{\ell}_{\theta^*}(X_i) + o_{P_0}(1). \quad (2.13)$$

Proof. The proof of this lemma is a more specific version of the proof found in Van der Vaart (1998) [18] on page 54. \square

Lemma 2.2 implies that for consistent maximum-likelihood estimators (sufficient conditions for consistency are given, for instance, in theorem 5.7 of van der Vaart (1998) [18]) the distribution of $\sqrt{n}(\hat{\theta}_n - \theta^*)$ has a normal limit with mean zero and covariance $V_{\theta^*}^{-1} P_0[\dot{\ell}_{\theta^*} \dot{\ell}_{\theta^*}^T] V_{\theta^*}^{-1}$. More important for present purposes, however, is the fact that according to (2.13), this sequence differs from Δ_{n,θ^*} only by a term of order $o_{P_0}(1)$. Since the total-variational distance $\|N_{\mu,\Sigma} - N_{\nu,\Sigma}\|$ is bounded by a multiple of $\|\mu - \nu\|$ as $(\mu \rightarrow \nu)$, the assertion of the Bernstein-Von Mises theorem can also be formulated with the sequence $\sqrt{n}(\hat{\theta}_n - \theta^*)$ as the locations for the normal limit sequence. Using the invariance of total-variation under rescaling and shifts, this leads to the conclusion that:

$$\sup_B \left| \Pi_n(\vartheta \in B \mid X_1, \dots, X_n) - N_{\hat{\theta}_n, n^{-1}V_{\theta^*}}(B) \right| \xrightarrow{P_0} 0,$$

which demonstrates the usual interpretation of the Bernstein-Von Mises theorem most clearly: the sequence of posteriors resembles more-and-more closely a sequence of ‘sharpening’ normal distributions centred at the maximum-likelihood estimators. More generally, any sequence of estimators satisfying (2.13) (*i.e.* any *best-regular* estimator sequence) may be used to centre the normal limit sequence on.

The conditions for lemma 2.2, which derive directly from a fairly general set of conditions for asymptotic normality in parametric M -estimation (see, theorem 5.23 in Van der Vaart (1998) [18]), are close to the conditions of the above Bernstein-Von Mises theorem. In the well-specified situation the Lipschitz condition (2.8) can be relaxed slightly and replaced by the condition of differentiability in quadratic mean.

It was noted in the introduction that the mismatch of the asymptotic covariance matrix $V_{\theta^*}^{-1} P_0[\dot{\ell}_{\theta^*} \dot{\ell}_{\theta^*}^T] V_{\theta^*}^{-1}$ and the limiting covariance matrix $V_{\theta^*}^{-1}$ in the Bernstein-Von Mises theorem causes that Bayesian credible sets are not confidence sets at the nominal level. The following example shows that both over- and under-covering may occur.

Example 2.1. Let P_θ be the normal distribution with mean θ and variance 1, and let the true distribution possess mean zero and variance $\sigma^2 > 0$. Then $\theta^* = 0$, $P_0 \dot{\ell}_{\theta^*}^2 = \sigma^2$ and $V_{\theta^*} = 1$. It follows that the radius of the $1 - \alpha$ -Bayesian credible set is z_α / \sqrt{n} , whereas a $1 - \alpha$ -confidence set around the mean has radius $z_\alpha \sigma / \sqrt{n}$. Depending on $\sigma^2 \leq 1$ or $\sigma^2 > 1$, the credible set can have coverage arbitrarily close to 0 or 1.

2.3. Asymptotic normality of point-estimators

Having discussed the posterior distributional limit, a natural question concerns the asymptotic properties of point-estimators derived from the posterior, like the posterior mean and median.

Based on the Bernstein-Von Mises assertion (2.4) alone, one sees that any functional $f : \mathcal{P} \mapsto \mathbb{R}$, continuous relative to the total-variational norm, when applied to the sequence of posterior laws, converges to f applied to the normal limit distribution. Another general consideration follows from a generic construction of point-estimates from posteriors and demonstrate that posterior consistency at rate δ_n implies frequentist consistency at rate δ_n .

Theorem 2.2. *Let X_1, \dots, X_n be distributed i.i.d.- P_0 and let $\Pi_n(\cdot | X_1, \dots, X_n)$ denote a sequence of posterior distributions on Θ that satisfies (2.3). Then there exist point-estimators $\hat{\theta}_n$ such that:*

$$\delta_n^{-1}(\hat{\theta}_n - \theta^*) = O_{P_0}(1), \quad (2.14)$$

i.e. $\hat{\theta}_n$ is consistent and converges to θ^* at rate δ_n .

Proof. Define $\hat{\theta}_n$ to be the center of a smallest ball that contains posterior mass at least $1/2$. Because the ball around θ^* of radius $\delta_n M_n$ contains posterior mass tending to 1, the radius of a smallest ball must be bounded by $\delta_n M_n$ and the smallest ball must intersect the ball of radius $\delta_n M_n$ around θ^* with probability tending to 1. This shows that $\|\hat{\theta}_n - \theta^*\| \leq 2\delta_n M_n$ with probability tending to one. \square

Consequently, frequentist restrictions and notions of asymptotic optimality have implications for the posterior too: in particular, frequentist bounds on the rate of convergence for a given problem apply to the posterior rate as well.

However, these general points are more appropriate in non-parametric context and the above existence theorem does not pertain to the most widely-used Bayesian point-estimators. Asymptotic normality of the posterior mean in a misspecified model has been analysed in Bunke and Milhaud (1998) [5]. We generalize their discussion and prove asymptotic normality and efficiency for a class of point-estimators defined by a general loss function, which includes the posterior mean and median.

Let $\ell : \mathbb{R}^k \rightarrow [0, \infty)$ be a loss-function with the following properties: ℓ is continuous and satisfies, for every $M > 0$,

$$\sup_{\|h\| \leq M} \ell(h) \leq \inf_{\|h\| > 2M} \ell(h),$$

with strict inequality for some M . Furthermore, we assume that ℓ is subpolynomial, i.e. for some $p > 0$,

$$\ell(h) \leq 1 + \|h\|^p. \quad (2.15)$$

Define the estimators $\hat{\theta}_n$ as the (near-)minimizers of

$$t \mapsto \int \ell(\sqrt{n}(t - \theta)) d\Pi_n(\theta | X_1, \dots, X_n).$$

The theorem below is the misspecified analog of theorem 10.8 in van der Vaart (1998) [18] and is based on general methods from M -estimation, in particular the argmax theorem (see, for example, corollary 5.58 in [18]).

Theorem 2.3. Assume that the model satisfies (2.1) for some $\theta^* \in \Theta$ and that the conditions of theorems 3.1 are satisfied. Let $\ell : \mathbb{R}^k \rightarrow [0, \infty)$ be a loss-function with the properties listed and assume that $\int \|\theta\|^p d\Pi(\theta) < \infty$. Then under P_0 , the sequence $\sqrt{n}(\hat{\theta}_n - \theta^*)$ converges weakly to the minimizer of,

$$t \mapsto Z(t) = \int \ell(t - h) dN_{X, V_{\theta^*}^{-1}}(h),$$

where $X \sim N(0, V_{\theta^*}^{-1} P_0[\dot{\ell}_{\theta^*} \dot{\ell}_{\theta^*}^T] V_{\theta^*}^{-1})$, provided that any two minimizers of this process coincide almost-surely. In particular, if the loss function is subconvex (e.g. $\ell(x) = \|x\|^2$ or $\ell(x) = \|x\|$, giving the posterior mean and median), then $\sqrt{n}(\hat{\theta}_n - \theta^*)$ converges weakly to X under P_0 .

Proof. The theorem can be proved along the same lines as theorem 10.8 in [18]. The main difference is in proving that, for any $M_n \rightarrow \infty$,

$$U_n := \int_{\|h\| > M_n} \|h\|^p d\Pi_n(h|X_1, \dots, X_n) \xrightarrow{P_0} 0. \quad (2.16)$$

Here, abusing notation, we write $d\Pi_n(h|X_1, \dots, X_n)$ to denote integrals relative to the posterior distribution of the local parameter $h = \sqrt{n}(\theta - \theta^*)$. Under misspecification a new proof is required, for which we extend the proof of theorem 3.1 below.

Once (2.16) is established, the proof continues as follows. The variable $\hat{h}_n = \sqrt{n}(\hat{\theta}_n - \theta^*)$ is the maximizer of the process $t \mapsto \int \ell(t - h) d\Pi_n(h|X_1, \dots, X_n)$. Reasoning exactly as in the proof of theorem 10.8, we see that $\hat{h}_n = O_{P_0}(1)$. Fix some compact set K and for given $M > 0$ define the processes

$$\begin{aligned} t \mapsto Z_{n,M}(t) &= \int_{\|h\| \leq M} \ell(t - h) d\Pi_n(h|X_1, \dots, X_n) \\ t \mapsto W_{n,M}(t) &= \int_{\|h\| \leq M} \ell(t - h) dN_{\Delta_n, V_{\theta^*}^{-1}}(h) \\ t \mapsto W_M(t) &= \int_{\|h\| \leq M} \ell(t - h) dN_{X, V_{\theta^*}^{-1}}(h) \end{aligned}$$

Since $\sup_{t \in K, \|h\| \leq M} \ell(t - h) < \infty$, $Z_{n,M} - W_{n,M} = o_{P_0}(1)$ in $\ell^\infty(K)$ by theorem 2.1. Since $\Delta_n \xrightarrow{P_0} X$, the continuous mapping theorem implies that $W_{n,M} \xrightarrow{P_0} W_M$ in $\ell^\infty(K)$. Since ℓ has subpolynomial tails, integrable with respect to $N_{X, V_{\theta^*}^{-1}}$, $W_M \xrightarrow{P_0} Z$ in $\ell^\infty(K)$ as $M \rightarrow \infty$. Thus $Z_{n,M} \xrightarrow{P_0} W_M$ in $\ell^\infty(K)$, for every $M > 0$, and $W_M \xrightarrow{P_0} Z$ as $M \rightarrow \infty$. We conclude that there exists a sequence $M_n \rightarrow \infty$ such that $Z_{n,M_n} \xrightarrow{P_0} Z$. The limit (2.16) implies that $Z_{n,M_n} - Z_n = o_{P_0}(1)$ in $\ell^\infty(K)$ and we conclude that $Z_n \xrightarrow{P_0} Z$ in $\ell^\infty(K)$. Due to the continuity of ℓ , $t \mapsto Z(t)$ is continuous almost surely. This, together with the assumed unicity of maxima of these sample paths, enables the argmax theorem (see, corollary 5.58 in [18]) and we conclude that $\hat{h}_n \xrightarrow{P_0} \hat{h}$, where \hat{h} is the minimizer of $Z(t)$.

For the proof of (2.16) we adopt the notation of theorem 3.1. The tests ω_n employed there can be taken nonrandomized without loss of generality (otherwise replace them for instance by $1_{\omega_n > 1/2}$) and then $U_n \omega_n$ tends to zero in probability by the only fact that ω_n does so. Thus (2.16) is proved once it is established that, with $\epsilon_n = M_n/\sqrt{n}$,

$$P_0^n(1 - \omega_n) 1_{\Omega \setminus \Xi_n} \int_{\epsilon_n \leq \|\theta - \theta^*\| < \epsilon} n^{p/2} \|\theta - \theta^*\|^p d\Pi_n(\theta \mid X_1, \dots, X_n) \rightarrow 0,$$

$$P_0^n(1 - \omega_n) 1_{\Omega \setminus \Omega_n} \int_{\|\theta - \theta^*\| \geq \epsilon} n^{p/2} \|\theta - \theta^*\|^p d\Pi_n(\theta \mid X_1, \dots, X_n) \rightarrow 0.$$

We can use bounds as in the proof of theorem 3.1, but instead of at (3.5) and (3.5) we arrive at the bounds

$$\frac{e^{n(a_n^2(1+C) - D\epsilon^2)}}{\Pi(B(a_n, \theta^*; P_0))} n^{p/2} \int \|\theta - \theta^*\|^p d\Pi(\theta),$$

$$K' e^{-\frac{1}{2}nD\epsilon_n^2} \sum_{j=1}^{\infty} n^{p/2} (j+1)^{d+p} \epsilon_n^p e^{-nD(j^2-1)\epsilon_n^2}.$$

These expressions tend to zero as before.

The last assertion of the theorem follows, because for a subconvex loss function the process Z is minimized uniquely by X , as a consequence of Anderson's lemma (see, for example, lemma 8.5 in [18]). \square

3. Rate of convergence

In a Bayesian context, the rate of convergence is defined as the maximal pace at which balls around the point of convergence can be shrunk to radius zero while still capturing a posterior mass that converges to one asymptotically. Current interest lies in the fact that the Bernstein-Von Mises theorem of the previous section formulates condition (2.3) (with $\delta_n = n^{-1/2}$),

$$\Pi_n(\|\vartheta - \theta^*\| \geq M_n/\sqrt{n} \mid X_1, \dots, X_n) \xrightarrow{P} 0,$$

for all $M_n \rightarrow \infty$. A convenient way of establishing the above is through the condition that suitable test sequences exist. As has been shown in a well-specified context in Ghosal *et al.* (2000) [7] and under misspecification in Kleijn and Van der Vaart (2003) [9], the most important requirement for convergence of the posterior at a certain rate is the existence of a test-sequence that separates the point of convergence from the complements of balls shrinking at said rate.

This is also the approach we follow here: we show that the sequence of posterior probabilities in the above display converges to zero in P_0 -probability if a test sequence exists that is suitable in the sense given above (see the proof of theorem 3.1). However, under the regularity conditions that were formulated to

establish local asymptotic normality under misspecification in the previous section, more can be said: not complements of shrinking balls, but fixed alternatives are to be suitably testable against P_0 , thus relaxing the testing condition considerably. Locally, the construction relies on score-tests to separate the point of convergence from complements of neighbourhoods shrinking at rate $1/\sqrt{n}$, using Bernstein's inequality to obtain exponential power. The tests for fixed alternatives are used to extend those local tests to the full model.

In this section we prove that a prior mass condition and suitable test sequences suffice to prove convergence at the rate required for the Bernstein-Von Mises theorem formulated in section 2. The theorem that begins the next subsection summarizes the conclusion. Throughout the section we consider the *i.i.d.* case, with notation as in section 2.2.

3.1. Posterior rate of convergence

With use of theorem 3.3, we formulate a theorem that ensures \sqrt{n} -rate of convergence for the posterior distributions of smooth, testable models with sufficient prior mass around the point of convergence. The testability condition is formulated using measures Q_θ , defined by,

$$Q_\theta(A) = P_0\left(\frac{p_\theta}{p_{\theta^*}}1_A\right),$$

for all $A \in \mathcal{A}$ and all $\theta \in \Theta$. Note that all Q_θ are dominated by P_0 and that $Q_{\theta^*} = P_0$. Also note that if the model is well-specified, then $P_{\theta^*} = P_0$ and $Q_\theta = P_\theta$ for all θ . Therefore the use of Q_θ instead of P_θ to formulate the testing condition is relevant only in the misspecified situation (see Kleijn and Van der Vaart (2006) [9] for more on this subject). The proof of theorem 3.1 makes use of Kullback-Leibler neighbourhoods of θ^* of the form:

$$B(\epsilon, \theta^*; P_0) = \left\{ \theta \in \Theta : -P_0 \log \frac{p_\theta}{p_{\theta^*}} \leq \epsilon^2, P_0 \left(\log \frac{p_\theta}{p_{\theta^*}} \right)^2 \leq \epsilon^2 \right\}, \quad (3.1)$$

for some $\epsilon > 0$.

Theorem 3.1. *Assume that the model satisfies the smoothness conditions of lemma 2.1, where in addition, it is required that $P_0(p_\theta/p_{\theta^*}) < \infty$ for all θ in a neighbourhood of θ^* and $P_0(e^{sm_{\theta^*}}) < \infty$ for some $s > 0$. Assume that the prior possesses a density that is continuous and positive in a neighbourhood of θ^* . Furthermore, assume that $P_0 \dot{\ell}_{\theta^*} \dot{\ell}_{\theta^*}^T$ is invertible and that for every $\epsilon > 0$ there exists a sequence of tests (ϕ_n) such that:*

$$P_0^n \phi_n \rightarrow 0, \quad \sup_{\{\theta: \|\theta - \theta^*\| \geq \epsilon\}} Q_\theta^n (1 - \phi_n) \rightarrow 0. \quad (3.2)$$

Then the posterior converges at rate $1/\sqrt{n}$, i.e. for every sequence (M_n) , $M_n \rightarrow \infty$:

$$\Pi_n(\theta \in \Theta : \|\theta - \theta^*\| \geq M_n/\sqrt{n} \mid X_1, X_2, \dots, X_n) \xrightarrow{P_0} 0.$$

Proof. Let (M_n) be given, and define the sequence (ϵ_n) by $\epsilon_n = M_n/\sqrt{n}$. According to theorem 3.3 there exists a sequence of tests (ω_n) and constants $D > 0$ and $\epsilon > 0$ such that (3.7) holds. We use these tests to split the P_0^n -expectation of the posterior measure as follows:

$$\begin{aligned} P_0^n \Pi(\theta : \|\theta - \theta^*\| \geq \epsilon_n \mid X_1, X_2, \dots, X_n) \\ \leq P_0^n \omega_n + P_0^n (1 - \omega_n) \Pi(\theta : \|\theta - \theta^*\| \geq \epsilon_n \mid X_1, X_2, \dots, X_n). \end{aligned}$$

The first term is of order $o(1)$ as $n \rightarrow \infty$ by (3.7). Given a constant $\epsilon > 0$ (to be specified later), the second term can be decomposed as:

$$\begin{aligned} P_0^n (1 - \omega_n) \Pi(\theta : \|\theta - \theta^*\| \geq \epsilon_n \mid X_1, X_2, \dots, X_n) \\ = P_0^n (1 - \omega_n) \Pi(\theta : \|\theta - \theta^*\| \geq \epsilon \mid X_1, X_2, \dots, X_n) \\ + P_0^n (1 - \omega_n) \Pi(\theta : \epsilon_n \leq \|\theta - \theta^*\| < \epsilon \mid X_1, X_2, \dots, X_n). \end{aligned} \quad (3.3)$$

Given two constants $M, M' > 0$ (also to be specified at a later stage), we define the sequences (a_n) , $a_n = M\sqrt{\log n/n}$ and (b_n) , $b_n = M'\epsilon_n$. Based on a_n and b_n , we define two sequences of events:

$$\begin{aligned} \Xi_n &= \left\{ \int_{\Theta} \prod_{i=1}^n \frac{p_{\theta}}{p_{\theta^*}}(X_i) d\Pi(\theta) \leq \Pi(B(a_n, \theta^*; P_0)) e^{-na_n^2(1+C)} \right\}, \\ \Omega_n &= \left\{ \int_{\Theta} \prod_{i=1}^n \frac{p_{\theta}}{p_{\theta^*}}(X_i) d\Pi(\theta) \leq \Pi(B(b_n, \theta^*; P_0)) e^{-nb_n^2(1+C)} \right\}. \end{aligned}$$

The sequence (Ξ_n) is used to split the first term on the *r.h.s.* of (3.3) and estimate it as follows:

$$\begin{aligned} P_0^n (1 - \omega_n) \Pi(\theta : \|\theta - \theta^*\| \geq \epsilon \mid X_1, X_2, \dots, X_n) \\ \leq P_0^n (\Xi_n) + P_0^n (1 - \omega_n) 1_{\Omega \setminus \Xi_n} \Pi(\theta : \|\theta - \theta^*\| \geq \epsilon \mid X_1, X_2, \dots, X_n). \end{aligned}$$

According to lemma 3.1, the first term is of order $o(1)$ as $n \rightarrow \infty$. The second term is estimated further with the use of lemmas 3.1, 3.2 and theorem 3.3: for some $C > 0$,

$$\begin{aligned} P_0^n (1 - \omega_n) 1_{\Omega \setminus \Xi_n} \Pi(\theta : \|\theta - \theta^*\| \geq \epsilon \mid X_1, X_2, \dots, X_n) \\ \leq \frac{e^{na_n^2(1+C)}}{\Pi(B(a_n, \theta^*; P_0))} \int_{\{\theta : \|\theta - \theta^*\| \geq \epsilon\}} Q_{\theta}^n (1 - \omega_n) d\Pi(\theta) \\ \leq \frac{e^{n(a_n^2(1+C) - D\epsilon^2)}}{\Pi(B(a_n, \theta^*; P_0))} \Pi(\theta : \|\theta - \theta^*\| \geq \epsilon). \end{aligned} \quad (3.4)$$

Note that $a_n^2(1+C) - D\epsilon^2 \leq -a_n^2(1+C)$ for large enough n , so that:

$$\frac{e^{n(a_n^2(1+C) - D\epsilon^2)}}{\Pi(B(a_n, \theta^*; P_0))} \leq K^{-1} e^{-na_n^2(1+C)} (a_n)^{-d} \leq \frac{1}{M^{d/2} K} (\log n)^{-d/2} n^{-M^2(1+C) + \frac{d}{2}},$$

for large enough n , using (3.6). A large enough choice for the constant M then ensures that the expression on the *l.h.s.* in the next-to-last display is of order $o(1)$ as $n \rightarrow \infty$.

The sequence (Ω_n) is used to split the second term on the *r.h.s.* of (3.3) after which we estimate it in a similar manner. Again the term that derives from 1_{Ω_n} is of order $o(1)$, and

$$\begin{aligned} & P_0^n(1-\omega_n) 1_{\Omega \setminus \Omega_n} \Pi(\theta : \epsilon_n \leq \|\theta - \theta^*\| < \epsilon \mid X_1, X_2, \dots, X_n) \\ & \leq \frac{e^{nb_n^2(1+C)}}{\Pi(B(b_n, \theta^*; P_0))} \sum_{j=1}^J \int_{A_{n,j}} Q_\theta^n(1-\omega_n) d\Pi(\theta), \end{aligned}$$

where we have split the domain of integration into spherical shells $A_{n,j}$, ($1 \leq j \leq J$, with J the smallest integer such that $(J+1)\epsilon_n > \epsilon$): $A_{n,j} = \{\theta : j\epsilon_n \leq \|\theta - \theta^*\| \leq ((j+1)\epsilon_n) \wedge \epsilon\}$. Applying theorem 3.3 to each of the shells separately, we obtain:

$$\begin{aligned} & P_0^n(1-\omega_n) 1_{\Omega \setminus \Omega_n} \Pi(\theta : \epsilon_n \leq \|\theta - \theta^*\| < \epsilon \mid X_1, X_2, \dots, X_n) \\ & = \sum_{j=1}^J e^{nb_n^2(1+C)} \sup_{\theta \in A_{n,j}} Q_\theta^n(1-\omega_n) \frac{\Pi(A_{n,j})}{\Pi(B(b_n, \theta^*; P_0))} \\ & \leq \sum_{j=1}^J e^{nb_n^2(1+C) - nDj^2\epsilon_n^2} \frac{\Pi\{\theta : \|\theta - \theta^*\| \leq (j+1)\epsilon_n\}}{\Pi(B(b_n, \theta^*; P_0))}. \end{aligned}$$

For a small enough ϵ and large enough n , the sets $\{\theta : \|\theta - \theta^*\| \leq (j+1)\epsilon_n\}$ all fall within the neighbourhood U of θ^* on which the prior density π is continuous. Hence π is uniformly bounded by a constant $R > 0$ and we see that: $\Pi\{\theta : \|\theta - \theta^*\| \leq (j+1)\epsilon_n\} \leq RV_d(j+1)^d\epsilon_n^d$, where V_d is the Lebesgue-volume of the d -dimensional ball of unit radius. Combining this with (3.6), there exists a constant $K' > 0$ such that, with $M' < \sqrt{D/2(1+C)}$:

$$\begin{aligned} & P_0^n(1-\omega_n) 1_{\Omega \setminus \Omega_n} \Pi(\theta : \epsilon_n \leq \|\theta - \theta^*\| < \epsilon \mid X_1, \dots, X_n) \\ & \leq K' e^{-\frac{1}{2}nD\epsilon_n^2} \sum_{j=1}^{\infty} (j+1)^d e^{-nD(j^2-1)\epsilon_n^2}, \end{aligned} \tag{3.5}$$

for large enough n . The series is convergent and we conclude that this term is also of order $o(1)$ as $n \rightarrow \infty$. \square

Consistent testability of the type (3.2) appears to be a weak requirement because the form of the tests is arbitrary. (It may be compared to “classical conditions” like (B3) in section 6.7, page 455, of [13], formulated in the well-specified case.) Consistent testability is of course one of Schwarz’ conditions for consistency ([16]) and appears to have been introduced in the context of the (well-specified) Bernstein-Von Mises theorem by Le Cam. To exemplify its power we show in the next theorem that the tests exist as soon as the parameter set is compact and the model is suitably continuous in the parameter.

Theorem 3.2. Assume that Θ is compact and that θ^* is a unique point of minimum of $\theta \mapsto -P_0 \log p_\theta$. Furthermore assume that $P_0(p_\theta/p_{\theta^*}) < \infty$ for all $\theta \in \Theta$ and that the map,

$$\theta \mapsto P_0\left(\frac{p_\theta}{p_{\theta_1}^s p_{\theta^*}^{1-s}}\right),$$

is continuous at θ_1 for every s in a left neighbourhood of 1, for every θ_1 . Then there exist tests ϕ_n satisfying (3.2). A sufficient condition is that for every $\theta_1 \in \Theta$ the maps $\theta \mapsto p_\theta/p_{\theta_1}$ and $\theta \mapsto p_\theta/p_{\theta^*}$ are continuous in $L_1(P_0)$ at $\theta = \theta_1$.

Proof. For given $\theta_1 \neq \theta^*$ consider the tests,

$$\phi_{n,\theta_1} = 1\{\mathbb{P}_n \log(p_0/q_{\theta_1}) < 0\}.$$

Because $\mathbb{P}_n \log(p_0/q_{\theta_1}) \rightarrow P_0 \log(p_0/q_{\theta_1})$ in P_0^n -probability by the law of large numbers, and $P_0 \log(p_0/q_{\theta_1}) = P_0 \log(p_{\theta^*}/p_{\theta_1}) > 0$ for $\theta_1 \neq \theta^*$ by the definition of θ^* we have that $P_0^n \phi_{n,\theta_1} \rightarrow 0$ as $n \rightarrow \infty$. By Markov's inequality we have that,

$$\begin{aligned} Q_\theta^n(1 - \phi_{n,\theta_1}) &= Q_\theta^n\left(e^{s n \mathbb{P}_n \log(p_0/q_{\theta_1})} > 1\right) \\ &\leq Q_\theta^n e^{s n \mathbb{P}_n \log(p_0/q_{\theta_1})} = \left(Q_\theta(p_0/q_{\theta_1})^s\right)^n = \rho(\theta_1, \theta, s)^n, \end{aligned}$$

for $\rho(\theta_1, \theta, s) = \int p_0^s q_{\theta_1}^{-s} q_\theta d\mu$. It is known from Kleijn and Van der Vaart (2006) [9] that the affinity $s \mapsto \rho(\theta_1, \theta_1, s)$ tends to $P_0(q_{\theta_1} > 0) = P_0(p_{\theta_1} > 0)$ as $s \uparrow 1$ and has derivative from the left equal to $P_0 \log(q_{\theta_1}/p_0) 1_{q_{\theta_1} > 0} = P_0 \log(p_{\theta_1}/p_{\theta^*}) 1_{p_{\theta_1} > 0}$ at $s = 1$. We have that either $P_0(p_{\theta_1} > 0) < 1$ or $P_0(p_{\theta_1} > 0) = 1$ and $P_0 \log(p_{\theta_1}/p_{\theta^*}) 1_{p_{\theta_1} > 0} = P_0 \log(p_{\theta_1}/p_{\theta^*}) < 0$ (or both). In all cases there exists $s_{\theta_1} < 1$ arbitrarily close to 1 such that $\rho(\theta_1, \theta_1, s_{\theta_1}) < 1$. By assumption the map $\theta \mapsto \rho(\theta_1, \theta, s_{\theta_1})$ is continuous at θ_1 . Therefore, for every θ_1 there exists an open neighbourhood G_{θ_1} such that,

$$r_{\theta_1} = \sup_{\theta \in G_{\theta_1}} \rho(\theta_1, \theta, s_{\theta_1}) < 1.$$

The set $\{\theta \in \Theta : \|\theta - \theta^*\| \geq \epsilon\}$ is compact and hence can be covered with finitely many sets of the type G_{θ_1} , say G_{θ_i} for $i = 1, \dots, k$. We now define

$$\phi_n = \max_{i=1, \dots, k} \phi_{n,\theta_i}.$$

This test satisfies

$$\begin{aligned} P_0^n \phi_n &\leq \sum_{i=1}^k P_0^n \phi_{n,\theta_i} \rightarrow 0, \\ Q_\theta^n(1 - \phi_n) &\leq \max_{i=1}^k Q_\theta^n(1 - \phi_{n,\theta_i}) \leq \max_{i=1}^k r_{\theta_i}^n \rightarrow 0, \end{aligned}$$

uniformly in $\theta \in \cup_{i=1}^k G_{\theta_i}$. Therefore the tests ϕ_n satisfy the requirements.

To prove the last assertion we write $\rho(\theta_1, \theta, s) = P_0(p_\theta/p_{\theta_1})^s (p_\theta/p_{\theta^*})^{1-s}$. Continuity of the maps $\theta \mapsto (p_\theta/p_{\theta_1})$ and $\theta \mapsto (p_\theta/p_{\theta^*})$ in $L_1(P_0)$ can be seen to imply the required continuity of the map $\theta \mapsto \rho(\theta_1, \theta, s)$. \square

Beyond compactness it appears impossible to give mere qualitative sufficient conditions, like continuity, for consistent testability. For “natural” parameterizations it ought to be true that distant parameters (outside a given compact) are the easy ones to test for (and a test designed for a given compact ought to be consistent even for points outside the compact), but this depends on the structure of the model. Alternatively, many models would allow a suitable compactification to which the preceding result can be applied, but we omit a discussion. The results in the next section allow to discard a “distant” part of the parameter space, after which the preceding results apply.

In the proof of theorem 3.1, lower bounds in probability on the denominators of posterior probabilities are needed, as provided by the following lemma.

Lemma 3.1. *For given $\epsilon > 0$ and $\theta^* \in \Theta$ such that $P_0 \log(p_0/p_{\theta^*}) < \infty$ define $B(\epsilon, \theta^*; P_0)$ by (3.1). Then for every $C > 0$ and probability measure Π on Θ :*

$$P_0^n \left(\int_{\Theta} \prod_{i=1}^n \frac{p_{\theta}}{p_{\theta^*}}(X_i) d\Pi(\theta) \leq \Pi(B(\epsilon, \theta^*; P_0)) e^{-n\epsilon^2(1+C)} \right) \leq \frac{1}{C^2 n \epsilon^2}.$$

Proof. This lemma can also be found as lemma 7.1 in Kleijn and Van der Vaart (2003) [9]. The proof is analogous to that of lemma 8.1 in Ghosal *et al.* (2000) [7]. \square

Moreover, the prior mass of the Kullback-Leibler neighbourhoods $B(\epsilon, \theta^*; P_0)$ can be lower-bounded if we make the regularity assumptions for the model used in section 2 and the assumption that the prior has a Lebesgue density that is well-behaved at θ^* .

Lemma 3.2. *Under the smoothness conditions of lemma 2.1 and assuming that the prior density π is continuous and strictly positive in θ^* , there exists a constant $K > 0$ such that the prior mass of the Kullback-Leibler neighbourhoods $B(\epsilon, \theta^*; P_0)$ satisfies:*

$$\Pi(B(\epsilon, \theta^*; P_0)) \geq K \epsilon^d, \quad (3.6)$$

for small enough $\epsilon > 0$.

Proof. As a result of the smoothness conditions, we have, for some constants $c_1, c_2 > 0$ and small enough $\|\theta - \theta^*\|$, $-P_0 \log(p_{\theta}/p_{\theta^*}) \leq c_1 \|\theta - \theta^*\|^2$, and $P_0(\log(p_{\theta}/p_{\theta^*}))^2 \leq c_2 \|\theta - \theta^*\|^2$. Defining $c = (1/c_1 \wedge 1/c_2)^{1/2}$, this implies that for small enough $\epsilon > 0$, $\{\theta \in \Theta : \|\theta - \theta^*\| \leq c\epsilon\} \subset B(\epsilon, \theta^*; P_0)$. Since the Lebesgue-density π of the prior is continuous and strictly positive in θ^* , we see that there exists a $\delta' > 0$ such that for all $0 < \delta \leq \delta'$: $\Pi(\theta \in \Theta : \|\theta - \theta^*\| \leq \delta) \geq \frac{1}{2} V_d \pi(\theta^*) \delta^d > 0$. Hence, for small enough ϵ , $c\epsilon \leq \delta'$ and we obtain (3.6) upon combination. \square

3.2. Suitable test sequences

In this subsection we prove that the existence of test sequences (under misspecification) of uniform exponential power for complements of shrinking balls

around θ^* versus P_0 (as needed in the proof of theorem 3.1), is guaranteed whenever asymptotically consistent test-sequences exist for complements of *fixed* balls around θ^* versus P_0 and the conditions of lemmas 2.1 and 3.4 are met. The following theorem is inspired by lemma 10.3 in Van der Vaart (1998) [18].

Theorem 3.3. *Assume that the conditions of lemma 2.1 are satisfied, where in addition, it is required that $P_0(p_\theta/p_{\theta^*}) < \infty$ for all θ in a neighbourhood of θ^* and $P_0(e^{sm_{\theta^*}}) < \infty$ for some $s > 0$. Furthermore, suppose that $P_0\dot{\ell}_{\theta^*}\dot{\ell}_{\theta^*}^T$ is invertible and for every $\epsilon > 0$ there exists a sequence of test functions (ϕ_n) , such that:*

$$P_0^n \phi_n \rightarrow 0, \quad \sup_{\{\theta: \|\theta - \theta^*\| \geq \epsilon\}} Q_\theta^n (1 - \phi_n) \rightarrow 0.$$

Then for every sequence (M_n) such that $M_n \rightarrow \infty$ there exists a sequence of tests (ω_n) such that for some constants $D > 0$, $\epsilon > 0$ and large enough n :

$$P_0^n \omega_n \rightarrow 0, \quad Q_\theta^n (1 - \omega_n) \leq e^{-nD(\|\theta - \theta^*\|^2 \wedge \epsilon^2)}, \quad (3.7)$$

for all $\theta \in \Theta$ such that $\|\theta - \theta^\| \geq M_n/\sqrt{n}$.*

Proof. Let (M_n) be given. We construct two sequences of tests: one sequence to test P_0 versus $\{Q_\theta : \theta \in \Theta_1\}$ with $\Theta_1 = \{\theta \in \Theta : M_n/\sqrt{n} \leq \|\theta - \theta^*\| \leq \epsilon\}$, and the other to test P_0 versus $\{Q_\theta : \theta \in \Theta_2\}$ with $\Theta_2 = \{\theta : \|\theta - \theta^*\| > \epsilon\}$, both uniformly with exponential power (for a suitable choice of ϵ). We combine these sequences to test P_0 versus $\{Q_\theta : \|\theta - \theta^*\| \geq M_n/\sqrt{n}\}$ uniformly with exponential power.

For the construction of the first sequence, a constant $L > 0$ is chosen to truncate the score-function component-wise (*i.e.* for all $1 \leq k \leq d$, $(\dot{\ell}_{\theta^*}^L)_k = 0$ if $|\dot{\ell}_{\theta^*}_k| \geq L$ and $(\dot{\ell}_{\theta^*}^L)_k = \dot{\ell}_{\theta^*}_k$ otherwise) and we define:

$$\omega_{1,n} = 1\{\|(\mathbb{P}_n - P_0)\dot{\ell}_{\theta^*}^L\| > \sqrt{M_n/n}\},$$

Because the function $\dot{\ell}_{\theta^*}$ is square-integrable, we can ensure that the matrices $P_0(\dot{\ell}_{\theta^*}\dot{\ell}_{\theta^*}^T)$, $P_0(\dot{\ell}_{\theta^*}(\dot{\ell}_{\theta^*}^L)^T)$ and $P_0(\dot{\ell}_{\theta^*}^L(\dot{\ell}_{\theta^*}^L)^T)$ are arbitrarily close (for instance in operator norm) by a sufficiently large choice for the constant L . We fix such an L throughout the proof.

By the central limit theorem $P_0^n \omega_{1,n} = P_0^n(\|\sqrt{n}(\mathbb{P}_n - P_0)\dot{\ell}_{\theta^*}^L\|^2 > M_n) \rightarrow 0$. Turning to $Q_\theta^n(1 - \omega_{1,n})$ for $\theta \in \Theta_1$, we note that for all θ :

$$\begin{aligned} Q_\theta^n\left(\|(\mathbb{P}_n - P_0)\dot{\ell}_{\theta^*}^L\| \leq \sqrt{M_n/n}\right) &= Q_\theta^n\left(\sup_{v \in S} v^T(\mathbb{P}_n - P_0)\dot{\ell}_{\theta^*}^L \leq \sqrt{M_n/n}\right) \\ &\leq \inf_{v \in S} Q_\theta^n\left(v^T(\mathbb{P}_n - P_0)\dot{\ell}_{\theta^*}^L \leq \sqrt{M_n/n}\right), \end{aligned}$$

where S is the sphere of unity in \mathbb{R}^d . With the choice $v = (\theta - \theta^*)/\|\theta - \theta^*\|$ as an upper bound for the *r.h.s.* in the above display, we note that:

$$\begin{aligned} Q_\theta^n\left((\theta - \theta^*)^T(\mathbb{P}_n - P_0)\dot{\ell}_{\theta^*}^L \leq \sqrt{\frac{M_n}{n}}\|\theta - \theta^*\|\right) \\ = Q_\theta^n\left((\theta^* - \theta)^T(\mathbb{P}_n - \tilde{Q}_\theta)\dot{\ell}_{\theta^*}^L \geq (\theta - \theta^*)^T(\tilde{Q}_\theta - \tilde{Q}_{\theta^*})\dot{\ell}_{\theta^*}^L - \sqrt{\frac{M_n}{n}}\|\theta - \theta^*\|\right), \end{aligned}$$

where we have used the notation (for all $\theta \in \Theta_1$ with small enough $\epsilon > 0$) $\tilde{Q}_\theta = \|Q_\theta\|^{-1}Q_\theta$ and the fact that $P_0 = Q_{\theta^*} = \tilde{Q}_{\theta^*}$. By straightforward manipulation, we find:

$$\begin{aligned} & (\theta - \theta^*)^T (\tilde{Q}_\theta - \tilde{Q}_{\theta^*}) \dot{\ell}_{\theta^*}^L \\ &= \frac{1}{P_0(p_\theta/p_{\theta^*})} (\theta - \theta^*)^T \left(P_0((p_\theta/p_{\theta^*} - 1) \dot{\ell}_{\theta^*}^L) + (1 - P_0(p_\theta/p_{\theta^*})) P_0 \dot{\ell}_{\theta^*}^L \right). \end{aligned}$$

In view of lemma 3.4 and conditions (2.8), (2.9), $(P_0(p_\theta/p_{\theta^*}) - 1)$ is of order $O(\|\theta - \theta^*\|^2)$ as $(\theta \rightarrow \theta^*)$, which means that if we approximate the above display up to order $o(\|\theta - \theta^*\|^2)$, we can limit attention on the *r.h.s.* to the first term in the last factor and equate the first factor to 1. Furthermore, using the differentiability of $\theta \mapsto \log(p_\theta/p_{\theta^*})$, condition (2.8) and lemma 3.4, we see that:

$$\begin{aligned} & P_0 \left\| \left(\frac{p_\theta}{p_{\theta^*}} - 1 - (\theta - \theta^*)^T \dot{\ell}_{\theta^*} \right) \dot{\ell}_{\theta^*}^L \right\| \\ & \leq P_0 \left\| \left(\frac{p_\theta}{p_{\theta^*}} - 1 - \log \frac{p_\theta}{p_{\theta^*}} \right) \dot{\ell}_{\theta^*}^L \right\| + P_0 \left\| \left(\log \frac{p_\theta}{p_{\theta^*}} - (\theta - \theta^*)^T \dot{\ell}_{\theta^*} \right) \dot{\ell}_{\theta^*}^L \right\|, \end{aligned}$$

which is $o(\|\theta - \theta^*\|)$. Also note that since $M_n \rightarrow \infty$ and for all $\theta \in \Theta_1$, $\|\theta - \theta^*\| \geq M_n/\sqrt{n}$, $-\|\theta - \theta^*\| \sqrt{M_n/n} \geq -\|\theta - \theta^*\|^2 (M_n)^{-1/2}$. Summarizing the above and combining with the remark made at the beginning of the proof concerning the choice of L , we find that for every $\delta > 0$, there exist choices of $\epsilon > 0$, $L > 0$ and $N \geq 1$ such that for all $n \geq N$ and all θ in Θ_1 :

$$\begin{aligned} & (\theta - \theta^*)^T (\tilde{Q}_\theta - \tilde{Q}_{\theta^*}) \dot{\ell}_{\theta^*}^L - \sqrt{M_n/n} \|\theta - \theta^*\| \\ & \geq (\theta - \theta^*)^T P_0 (\dot{\ell}_{\theta^*} \dot{\ell}_{\theta^*}^T) (\theta - \theta^*) - \delta \|\theta - \theta^*\|^2. \end{aligned}$$

We denote $\Delta(\theta) = (\theta - \theta^*)^T P_0 (\dot{\ell}_{\theta^*} \dot{\ell}_{\theta^*}^T) (\theta - \theta^*)$ and since $P_0 (\dot{\ell}_{\theta^*} \dot{\ell}_{\theta^*}^T)$ is strictly positive definite by assumption, its smallest eigenvalue c is greater than zero. Hence, $-\delta \|\theta - \theta^*\|^2 \geq -\delta/c \Delta(\theta)$. and there exists a constant $r(\delta)$ (depending only on the matrix $P_0 (\dot{\ell}_{\theta^*} \dot{\ell}_{\theta^*}^T)$ and with the property that $r(\delta) \rightarrow 1$ if $\delta \rightarrow 0$) such that:

$$Q_\theta^n (1 - \omega_{1,n}) \leq Q_\theta^n \left((\theta^* - \theta)^T (\mathbb{P}_n - \tilde{Q}_\theta) \dot{\ell}_{\theta^*}^L \geq r(\delta) \Delta(\theta) \right),$$

for small enough ϵ , large enough L and large enough n , demonstrating that the type-II error is bounded above by the (unnormalized) tail probability $Q_\theta^n (\bar{W}_n \geq r(\delta) \Delta(\theta))$ of the mean of the variables $W_i = (\theta^* - \theta)^T (\dot{\ell}_{\theta^*}^L(X_i) - \tilde{Q}_\theta \dot{\ell}_{\theta^*}^L)$, ($1 \leq i \leq n$), so that $\tilde{Q}_\theta W_i = 0$. The random variables W_i are independent and bounded since:

$$|W_i| \leq \|\theta - \theta^*\| (\|\dot{\ell}_{\theta^*}^L(X_i)\| + \|\tilde{Q}_\theta \dot{\ell}_{\theta^*}^L\|) \leq 2L\sqrt{d} \|\theta - \theta^*\|.$$

The variance of W_i under \tilde{Q}_θ is expressed as follows:

$$\text{Var}_{\tilde{Q}_\theta} W_i = (\theta - \theta^*)^T \left[\tilde{Q}_\theta (\dot{\ell}_{\theta^*}^L (\dot{\ell}_{\theta^*}^L)^T) - \tilde{Q}_\theta \dot{\ell}_{\theta^*}^L \tilde{Q}_\theta (\dot{\ell}_{\theta^*}^L)^T \right] (\theta - \theta^*).$$

Using that $P_0 \dot{\ell}_{\theta^*} = 0$ (see (2.11)), the above can be estimated like before, with the result that there exists a constant $s(\delta)$ (depending only on (the largest eigenvalue of) the matrix $P_0 \dot{\ell}_{\theta^*} \dot{\ell}_{\theta^*}^T$ and with the property that $s(\delta) \rightarrow 1$ as $\delta \rightarrow 0$) such that:

$$\text{Var}_{\tilde{Q}_\theta}(W_i) \leq s(\delta) \Delta(\theta),$$

for small enough ϵ and large enough L . We apply Bernstein's inequality (see, for instance, Pollard (1984) [15], pp. 192–193) to obtain:

$$\begin{aligned} Q_\theta^n(1 - \omega_{1,n}) &= \|Q_\theta\|^n \tilde{Q}_\theta^n(W_1 + \dots + W_n \geq nr(\delta) \Delta(\theta)) \\ &\leq \|Q_\theta\|^n \exp\left(-\frac{1}{2} \frac{r(\delta)^2 n \Delta(\theta)}{s(\delta) + \frac{3}{2} L \sqrt{d} \|\theta - \theta^*\| r(\delta)}\right). \end{aligned} \quad (3.8)$$

The factor $t(\delta) = r(\delta)^2 (s(\delta) + \frac{3}{2} L \sqrt{d} \|\theta - \theta^*\| r(\delta))^{-1}$ lies arbitrarily close to 1 for sufficiently small choices of δ and ϵ . As for the n -th power of the norm of Q_θ , we use lemma 3.4, (2.8) and (2.9) to estimate the norm of Q_θ as follows:

$$\begin{aligned} \|Q_\theta\| &= 1 + P_0 \log \frac{p_\theta}{p_{\theta^*}} + \frac{1}{2} P_0 \left(\log \frac{p_\theta}{p_{\theta^*}} \right)^2 + o(\|\theta - \theta^*\|^2) \\ &\leq 1 + P_0 \log \frac{p_\theta}{p_{\theta^*}} + \frac{1}{2} (\theta - \theta^*)^T P_0 (\dot{\ell}_{\theta^*} \dot{\ell}_{\theta^*}^T) (\theta - \theta^*) + o(\|\theta - \theta^*\|^2) \\ &\leq 1 - \frac{1}{2} (\theta - \theta^*)^T V_{\theta^*} (\theta - \theta^*) + \frac{1}{2} u(\delta) \Delta(\theta), \end{aligned} \quad (3.9)$$

for some constant $u(\delta)$ such that $u(\delta) \rightarrow 1$ if $\delta \rightarrow 0$. Because $1 + x \leq e^x$ for all $x \in \mathbb{R}$, we obtain, for sufficiently small $\|\theta - \theta^*\|$:

$$Q_\theta^n(1 - \omega_{1,n}) \leq \exp\left(-\frac{n}{2} (\theta - \theta^*)^T V_{\theta^*} (\theta - \theta^*) + \frac{n}{2} (u(\delta) - t(\delta)) \Delta(\theta)\right). \quad (3.10)$$

Note that $u(\delta) - t(\delta) \rightarrow 0$ as $\delta \rightarrow 0$ and $\Delta(\theta)$ is upper bounded by a multiple of $\|\theta - \theta^*\|^2$. Since V_{θ^*} is assumed to be invertible, we conclude that there exists a constant $C > 0$ such that for large enough L , small enough $\epsilon > 0$ and large enough n :

$$Q_\theta^n(1 - \omega_{1,n}) \leq e^{-Cn \|\theta - \theta^*\|^2}. \quad (3.11)$$

Concerning the range $\|\theta - \theta^*\| > \epsilon$, an asymptotically consistent test-sequence of P_0 versus Q_θ exists by assumption, what remains is the exponential power; the proof of lemma 3.3 demonstrates the existence of a sequence of tests $(\omega_{2,n})$ such that (3.12) holds. The sequence (ψ_n) is defined as the maximum of the two sequences defined above: $\psi_n = \omega_{1,n} \vee \omega_{2,n}$ for all $n \geq 1$, in which case $P_0^n \psi_n \leq P_0^n \omega_{1,n} + P_0^n \omega_{2,n} \rightarrow 0$ and:

$$\begin{aligned} \sup_{\theta \in A_n} Q_\theta^n(1 - \psi_n) &= \sup_{\theta \in \Theta_1} Q_\theta^n(1 - \psi_n) \vee \sup_{\theta \in \Theta_2} Q_\theta^n(1 - \psi_n) \\ &\leq \sup_{\theta \in \Theta_1} Q_\theta^n(1 - \omega_{1,n}) \vee \sup_{\theta \in \Theta_2} Q_\theta^n(1 - \omega_{2,n}). \end{aligned}$$

Combination of the bounds found in (3.11) and (3.12) and a suitable choice for the constant $D > 0$ lead to (3.7). \square

The following lemma shows that for a sequence of tests that separates P_0 from a fixed model subset V , there exists a exponentially powerful version without further conditions. Note that this lemma holds in non-parametric and parametric situations alike.

Lemma 3.3. *Suppose that for given measurable subset V of Θ , there exists a sequence of tests (ϕ_n) such that:*

$$P_0^n \phi_n \rightarrow 0, \quad \sup_{\theta \in V} Q_\theta^n (1 - \phi_n) \rightarrow 0.$$

Then there exists a sequence of tests (ω_n) and strictly positive constants C, D such that:

$$P_0^n \omega_n \leq e^{-nC}, \quad \sup_{\theta \in V} Q_\theta^n (1 - \omega_n) \leq e^{-nD}. \quad (3.12)$$

Proof. For given $0 < \zeta < 1$, we split the model subset V in two disjoint parts V_1 and V_2 defined by $V_1 = \{\theta \in V : \|Q_\theta\| \geq 1 - \zeta\}$, $V_2 = \{\theta \in V : \|Q_\theta\| < 1 - \zeta\}$. Note that for every test-sequence (ω_n) ,

$$\sup_{\theta \in V} Q_\theta^n (1 - \omega_n) \leq \sup_{\theta \in V_1} Q_\theta^n (1 - \omega_n) \vee (1 - \zeta)^n. \quad (3.13)$$

Let $\delta > 0$ be given. By assumption there exists an $N \geq 1$ such that for all $n \geq N + 1$, $P_0^n \phi_n \leq \delta$ and $\sup_{\theta \in V} Q_\theta^n (1 - \phi_n) \leq \delta$. Every $n \geq N + 1$ can be written as an m -fold multiple of N ($m \geq 1$) plus a remainder $1 \leq r \leq N$: $n = mN + r$. Given $n \geq N$, we divide the sample X_1, X_2, \dots, X_n into $(m - 1)$ groups of N consecutive X 's and a group of $N + r$ X 's and apply ϕ_N to the first $(m - 1)$ groups and ϕ_{N+r} to the last group, to obtain:

$$\begin{aligned} Y_{1,n} &= \phi_N(X_1, X_2, \dots, X_N), \\ Y_{2,n} &= \phi_N(X_{N+1}, X_{N+2}, \dots, X_{2N}), \\ &\vdots \\ Y_{m-1,n} &= \phi_N(X_{(m-2)N+1}, X_{(m-2)N+2}, \dots, X_{(m-1)N}), \\ Y_{m,n} &= \phi_{N+r}(X_{(m-1)N+1}, X_{(m-1)N+2}, \dots, X_{mN+r}), \end{aligned}$$

which are bounded, $0 \leq Y_{j,n} \leq 1$ for all $1 \leq j \leq m$ and $n \geq 1$. From that we define the test-statistic $\bar{Y}_{m,n} = (1/m)(Y_{1,n} + \dots + Y_{m,n})$ and the test-function $\omega_n = 1\{\bar{Y}_{m,n} \geq \eta\}$, based on a critical value $\eta > 0$ to be chosen at a later stage. The P_0^n -expectation of the test-function can be bounded as follows:

$$\begin{aligned} P_0^n \omega_n &= P_0^n (Y_{1,n} + \dots + Y_{m,n} \geq m\eta) \\ &= P_0^n \left(Z_{1,n} + \dots + Z_{m,n} \geq m\eta - \sum_{j=1}^{m-1} P_0^n Y_{j,n} - P_0^{N+r} Y_{m,n} \right) \\ &\leq P_0^n (Z_{1,n} + \dots + Z_{m,n} \geq m(\eta - \delta)), \end{aligned}$$

where $Z_{j,n} = Y_{j,n} - P_0^n Y_{j,n}$ for all $1 \leq j \leq m-1$ and $Z_{m,n} = Y_{m,n} - P_0^{N+r} Y_{m,n}$. Furthermore, the variables $Z_{j,n}$ are bounded $a_j \leq Z_{j,n} \leq b_j$ where $b_j - a_j = 1$. Imposing $\eta > \delta$ we may use Hoeffding's inequality to conclude that:

$$P_0^n \omega_n \leq e^{-2m(\eta-\delta)^2}. \quad (3.14)$$

A similar bound can be derived for $Q_\theta(1 - \omega_n)$ as follows. First we note that:

$$Q_\theta^n(1 - \omega_n) = Q_\theta^n \left(Z_{1,n} + \dots + Z_{m,n} > -m\eta + \sum_{j=1}^{m-1} Q_\theta^N Y_{j,n} + Q_\theta^{N+r} Y_{m,n} \right),$$

where, in this case, we have used the following definitions for the variables $Z_{j,n}$:

$$Z_{j,n} = -Y_{j,n} + Q_\theta^N Y_{j,n}, \quad Z_{m,n} = -Y_{m,n} + Q_\theta^{N+r} Y_{m,n},$$

for $1 \leq j \leq m-1$. We see that $a_j \leq Z_{j,n} \leq b_j$ with $b_j - a_j = 1$. Choosing $\zeta \leq 1 - (4\delta)^{1/N}$ (for small enough $\delta > 0$) and η between δ and 2δ , we see that for all $\theta \in V_1$:

$$\sum_{j=1}^{m-1} Q_\theta^N Y_{j,n} + Q_\theta^{N+r} Y_{m,n} - m\eta \geq m((1 - \zeta)^N - 3\delta) \geq m\delta > 0.$$

Hoeffding's inequality then provides the bound,

$$Q_\theta^n(1 - \omega_n) \leq \exp\left(-\frac{1}{2}m(\|Q_\theta\| - 3\delta)^2 + m \log \|Q_\theta\|^N\right).$$

In the case that $\|Q_\theta\| < 1$, we see that $Q_\theta^n(1 - \omega_n) \leq e^{-\frac{1}{2}m\delta^2}$. In the case that $\|Q_\theta\| \geq 1$, we use the identity $\log q \leq q - 1$ and the fact that $-\frac{1}{2}(q - 3\delta)^2 + (q - 1)$ has no zeroes for $q \in [1, \infty)$ if we choose $\delta < 1/6$, to conclude that the exponent is negative and bounded away from 0: $Q_\theta^n(1 - \omega_n) \leq e^{-mc}$ for some $c > 0$. Combining the two bounds leads to the assertion, if we notice that $m = (n - r)/N \geq n/N - 1$, absorbing eventual constants multiplying the exponential factor in (3.7) by a lower choice of D . \square

The following lemma is used in the proof of theorem 3.3 to control the behaviour of $\|Q_\theta\|$ in neighbourhoods of θ^* .

Lemma 3.4. *Assume that $P_0(p_\theta/p_{\theta^*})$ and $-P_0 \log(p_\theta/p_0)$ are finite for all θ in a neighbourhood U' of θ^* . Furthermore, assume that there exist a measurable function m such that,*

$$\left| \log \frac{p_\theta}{p_{\theta^*}} \right| \leq m\|\theta - \theta^*\|, \quad (P_0 - a.s.). \quad (3.15)$$

for all $\theta \in U'$ and such that $P_0(e^{sm}) < \infty$ for some $s > 0$. Then,

$$P_0 \left| \frac{p_\theta}{p_{\theta^*}} - 1 - \log \frac{p_\theta}{p_{\theta^*}} - \frac{1}{2} \left(\log \frac{p_\theta}{p_{\theta^*}} \right)^2 \right| = o(\|\theta - \theta^*\|^2).$$

Proof. The function $R(x)$ defined by $e^x = 1 + x + \frac{1}{2}x^2 + x^2 R(x)$ increases from $-\frac{1}{2}$ in the limit ($x \rightarrow -\infty$) to ∞ as ($x \rightarrow \infty$), with $R(x) \rightarrow R(0) = 0$ if ($x \rightarrow 0$). We also have $|R(-x)| \leq R(x) \leq e^x/x^2$ for all $x > 0$. The *l.h.s.* of the assertion of the lemma can be written as

$$P_0 \left(\log \frac{p_\theta}{p_{\theta^*}} \right)^2 \left| R \left(\log \frac{p_\theta}{p_{\theta^*}} \right) \right| \leq \|\theta - \theta^*\|^2 P_0 (m^2 R(m\|\theta - \theta^*\|)).$$

The expectation on the *r.h.s.* of the above display is bounded by $P_0 m_\theta^2 R(\epsilon m_\theta)$ if $\|\theta - \theta^*\| \leq \epsilon$. The functions $m^2 R(\epsilon m)$ are dominated by e^{sm} for sufficiently small ϵ and converge pointwise to $m^2 R(0) = 0$ as $\epsilon \downarrow 0$. The lemma then follows from the dominated convergence theorem. \square

4. Consistency and testability

The conditions for the theorems concerning rates of convergence and limiting behaviour of the posterior distribution discussed in the previous sections include several requirements on the model involving the true distribution P_0 . Depending on the specific model and true distribution, these requirements may be rather stringent, disqualifying for instance models in which $-P_0 \log p_\theta/p_{\theta^*} = \infty$ for θ in neighbourhoods of θ^* . To drop this kind of condition from the formulation and nevertheless maintain the current proof(s), we have to find other means to deal with ‘undesirable’ subsets of the model. In this section we show that if Kullback-Leibler neighbourhoods of the point of convergence receive enough prior mass and asymptotically consistent uniform tests for P_0 versus such subsets exist, they can be excluded from the model beforehand. As a special case, we derive a misspecified version of Schwartz’ consistency theorem (see Schwartz (1965) [16]). Results presented in this section hold for the parametric models considered in previous sections, but are also valid in non-parametric situations.

4.1. Exclusion of testable model subsets

We start by formulating and proving the lemma announced above, in its most general form.

Lemma 4.1. *Let $V \subset \Theta$ be a (measurable) subset of the model Θ . Assume that for some $\epsilon > 0$:*

$$\Pi \left(\theta \in \Theta : -P_0 \log \frac{p_\theta}{p_{\theta^*}} \leq \epsilon \right) > 0, \quad (4.1)$$

and there exist constants $\gamma > 0$, $\beta > \epsilon$ and a sequence (ϕ_n) of test-functions such that:

$$P_0^n \phi_n \leq e^{-n\gamma}, \quad \sup_{\theta \in V} Q_\theta^n (1 - \phi_n) \leq e^{-n\beta}, \quad (4.2)$$

for large enough $n \geq 1$. Then $\Pi(V|X_1, X_2, \dots, X_n) \rightarrow 0$, P_0 - a.s.

Proof. Due to the first inequality of (4.2), Markov's inequality and the first Borel-Cantelli lemma suffice to show that $\phi_n \rightarrow 0$, P_0 -almost-surely. We split the posterior measure of V with the test function ϕ_n and calculate the limes superior,

$$\limsup_{n \rightarrow \infty} \Pi(V|X_1, X_2, \dots, X_n) = \limsup_{n \rightarrow \infty} \Pi(V|X_1, X_2, \dots, X_n)(1 - \phi_n),$$

P_0 -almost-surely. Next, consider the subset $K_\epsilon = \{\theta \in \Theta : -P_0 \log(p_\theta/p_{\theta^*}) \leq \epsilon\}$. For every $\theta \in K_\epsilon$, the strong law of large numbers says that:

$$\left| \mathbb{P}_n \log \frac{p_\theta}{p_{\theta^*}} - P_0 \log \frac{p_\theta}{p_{\theta^*}} \right| \rightarrow 0,$$

P_0 -almost-surely. Hence for every $\alpha > \epsilon$ and all $\theta \in K_\epsilon$, there exists an $N \geq 1$ such that for all $n \geq N$, $\prod_{i=1}^n (p_\theta/p_{\theta^*})(X_i) \geq e^{-n\alpha}$, P_0^n -almost-surely. Therefore,

$$\begin{aligned} \liminf_{n \rightarrow \infty} e^{n\alpha} \int_{\Theta} \prod_{i=1}^n \frac{p_\theta}{p_{\theta^*}}(X_i) d\Pi(\theta) &\geq \liminf_{n \rightarrow \infty} e^{n\alpha} \int_{K_\epsilon} \prod_{i=1}^n \frac{p_\theta}{p_{\theta^*}}(X_i) d\Pi(\theta) \\ &\geq \int_{K_\epsilon} \liminf_{n \rightarrow \infty} e^{n\alpha} \prod_{i=1}^n \frac{p_\theta}{p_{\theta^*}}(X_i) d\Pi(\theta) \geq \Pi(K_\epsilon), \end{aligned}$$

by Fatou's lemma. Since by assumption, $\Pi(K_\epsilon) > 0$ we see that:

$$\begin{aligned} &\limsup_{n \rightarrow \infty} \Pi(V|X_1, X_2, \dots, X_n)(1 - \phi_n) \\ &\leq \frac{\limsup_{n \rightarrow \infty} e^{n\alpha} \int_V \prod_{i=1}^n \frac{p_\theta}{p_{\theta^*}}(X_i) (1 - \phi_n)(X_1, X_2, \dots, X_n) d\Pi(\theta)}{\liminf_{n \rightarrow \infty} e^{n\alpha} \int_{\Theta} \prod_{i=1}^n \frac{p_\theta}{p_{\theta^*}}(X_i) d\Pi(\theta)} \quad (4.3) \\ &\leq \frac{1}{\Pi(K_\epsilon)} \limsup_{n \rightarrow \infty} f_n(X_1, X_2, \dots, X_n), \end{aligned}$$

where we use the notation $f_n : \mathcal{X}^n \rightarrow \mathbb{R}$:

$$f_n(X_1, X_2, \dots, X_n) = e^{n\alpha} \int_V \prod_{i=1}^n \frac{p_\theta}{p_{\theta^*}}(X_i) (1 - \phi_n)(X_1, X_2, \dots, X_n) d\Pi(\theta).$$

Fubini's theorem and the assumption that the test-sequence is uniformly exponential, guarantee that for large enough n , $P_0^n f_n \leq e^{-n(\beta-\alpha)}$. Markov's inequality can then be used to show that:

$$P_0^\infty(f_n > e^{-\frac{n}{2}(\beta-\epsilon)}) \leq e^{n(\alpha-\frac{1}{2}(\beta+\epsilon))}.$$

Since $\beta > \epsilon$, we can choose α such that $\epsilon < \alpha < \frac{1}{2}(\beta + \epsilon)$ so that the series $\sum_{n=1}^\infty P_0^\infty(f_n > \exp -\frac{n}{2}(\beta - \epsilon))$ converges. The first Borel-Cantelli lemma then leads to the conclusion that:

$$P_0^\infty\left(\bigcap_{N=1}^\infty \bigcup_{n \geq N} \{f_n > e^{-\frac{n}{2}(\beta-\epsilon)}\}\right) = P_0^\infty\left(\limsup_{n \rightarrow \infty} (f_n - e^{-\frac{n}{2}(\beta-\epsilon)}) > 0\right) = 0.$$

Since $f_n \geq 0$, we see that $f_n \rightarrow 0$, P_0 -almost-surely, to conclude the proof. \square

In many situations, (4.1) is satisfied for every $\epsilon > 0$. In that case the construction of uniform exponentially powerful tests from asymptotically consistent tests (as demonstrated in the proof of lemma 3.3) can be used to fulfill (4.2) under the condition that an asymptotically consistent uniform test-sequence exists.

Corollary 4.1. *Let $V \subset \Theta$ be a (measurable) subset of the model Θ . Assume that for all $\epsilon > 0$ (4.1) is satisfied and that there exists a test-sequence (ϕ_n) such that:*

$$P_0^n \phi_n \rightarrow 0, \quad \sup_{\theta \in V} Q_\theta^n (1 - \phi_n) \rightarrow 0. \quad (4.4)$$

Then $\Pi(V|X_1, X_2, \dots, X_n) \rightarrow 0$, P_0 -almost-surely.

In this corollary form, the usefulness of lemma 4.1 is most apparent. All subsets V of the model that can be distinguished from P_0 based on a characteristic property (formalised by the test functions above) in a uniform manner (c.f. (4.4)) may be discarded from proofs like that of theorem 3.1. Hence the properties assumed in the statement of (for instance) theorem 3.1, can be left out as conditions if a suitable test sequence exist.

Whether or not a suitable test sequence can be found depends on the particular model and true distribution in question and little can be said in any generality. The likelihood ratio test is one possibility. The following lemma is comparable to the classical condition, as in [13].

Lemma 4.2. *Let $V \in \Theta$ be a (measurable) subset of the model Θ . Assume that for all $\epsilon > 0$ (4.1) is satisfied and suppose that there exists a sequence (M_n) of positive numbers such that $M_n \rightarrow \infty$ and*

$$P_0^n \left(\inf_{\theta \in V} -\mathbb{P}_n \log \frac{p_\theta}{p_{\theta^*}} < \frac{1}{n} M_n \right) \rightarrow 0. \quad (4.5)$$

Then $\Pi(V|X_1, X_2, \dots, X_n) \rightarrow 0$, P_0 -almost-surely.

Proof. Define the sequence of test functions:

$$\psi_n = 1 \left\{ \inf_{\theta \in V} -\mathbb{P}_n \log \frac{p_\theta}{p_{\theta^*}} < \frac{1}{n} M_n \right\}.$$

According to assumption (4.5), $P_0^n \psi_n \rightarrow 0$. Let $\theta \in V$ be given.

$$\begin{aligned} Q_\theta^n (1 - \psi_n) &= P_0^n \left(\frac{dP_\theta^n}{dP_{\theta^*}^n} (1 - \psi) \right) = P_0^n \left(\frac{dP_\theta^n}{dP_{\theta^*}^n} 1 \left\{ \sup_{\theta \in V} \log \frac{dP_\theta^n}{dP_{\theta^*}^n} \leq -M_n \right\} \right) \\ &\leq e^{-M_n} P_0^n \left(\sup_{\theta \in V} \log \frac{dP_\theta^n}{dP_{\theta^*}^n} \leq -M_n \right) \leq e^{-M_n} \rightarrow 0. \end{aligned}$$

Since M_n does not depend on θ , convergence to 0 is uniform over V . Corollary 4.1 then gives the assertion. \square

Finally, we note that lemma 4.1 can also be used to prove consistency. Presently, we do not assume the existence of a *unique* minimizer of the Kullback-Leibler divergence; we define Θ^* to be the set of points in the model at minimal

Kullback-Leibler divergence with respect to the true distribution P_0 , $\Theta^* = \{\theta \in \Theta : -P_0 \log(p_\theta/p_0) = \inf_{\Theta} -P_0 \log(p_\theta/p_0)\}$, and we consider the posterior probability of this set under the conditions of corollary 4.1. We write $d(\theta, \Theta^*)$ for the infimum of $\|\theta - \theta^*\|$ over $\theta^* \in \Theta^*$.

Corollary 4.2. (*Schwartz consistency*) Assume that for all $\epsilon > 0$ (4.1) is satisfied and that for all $\eta > 0$ there exists a test-sequence (ϕ_n) such that:

$$P_0^n \phi_n \rightarrow 0, \quad \sup_{\theta: d(\theta, \Theta^*) > \eta} Q_\theta^n(1 - \phi_n) \rightarrow 0.$$

Then $\Pi(d(\theta, \Theta^*) > \eta | X_1, X_2, \dots, X_n) \rightarrow 1$, P_0 -almost-surely, for every $\eta > 0$.

5. Appendix: Technical lemmas

The first lemma used in the proof of theorem 2.1 shows that given two sequences of probability measures, a sequence of balls that grows fast enough can be used conditionally to calculate the difference in total-variational distance, even when the sequences consist of random measures.

Lemma 5.1. Let (Π_n) and (Φ_n) be two sequences of random probability measures on \mathbb{R}^d . Let (K_n) be a sequence of subsets of \mathbb{R}^d such that,

$$\Pi_n(\mathbb{R}^d \setminus K_n) \xrightarrow{P_0} 0, \quad \Phi_n(\mathbb{R}^d \setminus K_n) \xrightarrow{P_0} 0. \quad (5.1)$$

Then,

$$\|\Pi_n - \Phi_n\| - \|\Pi_n^{K_n} - \Phi_n^{K_n}\| \xrightarrow{P_0} 0. \quad (5.2)$$

Proof. Let K , a measurable subset of \mathbb{R}^d and $n \geq 1$ be given and assume that $\Pi_n(K) > 0$ and $\Phi_n(K) > 0$. Then for any measurable $B \subset \mathbb{R}^d$ we have:

$$|\Pi_n(B) - \Pi_n^K(B)| \leq 2\Pi_n(\mathbb{R}^d \setminus K).$$

and hence also:

$$\left| (\Pi_n(B) - \Pi_n^K(B)) - (\Phi_n(B) - \Phi_n^K(B)) \right| \leq 2(\Pi_n(\mathbb{R}^d \setminus K) + \Phi_n(\mathbb{R}^d \setminus K)). \quad (5.3)$$

As a result of the triangle inequality, we then find that the difference in total-variation distances between Π_n and Φ_n on the one hand and Π_n^K and Φ_n^K on the other is bounded above by the expression on the right in the above display (which is independent of B).

Define A_n, B_n to be the events that $\Pi_n(K_n) > 0, \Phi_n(K_n) > 0$ respectively. On $\Xi_n = A_n \cap B_n$, $\Pi_n^{K_n}$ and $\Phi_n^{K_n}$ are well-defined probability measures. Assumption (5.1) guarantees that $P_0^n(\Xi_n)$ converges to 1. Restricting attention to the event Ξ_n in the above upon substitution of the sequence (K_n) and using (5.1) for the limit of (5.3) we find (5.2), where it is understood that the conditional probabilities on the l.h.s. are well-defined with probability growing to 1. \square

The second lemma demonstrates that the sequence of normals satisfies the condition of lemma 5.1 when the sequence of centre points Δ_{n,θ^*} is uniformly tight.

Lemma 5.2. *Let K_n be a sequence of balls centred on the origin with radii $M_n \rightarrow \infty$. Let (Φ_n) be a sequence of normal distributions (with fixed covariance matrix V) located respectively at the (random) points $(\Delta_n) \subset \mathbb{R}^d$. If the sequence Δ_n is uniformly tight, then $\Phi_n(\mathbb{R}^d \setminus K_n) = N_{\Delta_n, V}(\mathbb{R}^d \setminus K_n) \xrightarrow{P_0} 0$.*

Proof. Let $\delta > 0$ be given. Uniform tightness of the sequence (Δ_n) implies the existence of a constant $L > 0$ such that:

$$\sup_{n \geq 1} P_0^n(\|\Delta_n\| \geq L) \leq \delta.$$

For all $n \geq 1$, call $A_n = \{\|\Delta_n\| \geq L\}$. Let $\mu \in \mathbb{R}^d$ be given. Since $N(\mu, V)$ is tight, there exists for every given $\epsilon > 0$ a constant L' such that $N_{\mu, V}(B(\mu, L')) \geq 1 - \epsilon$ (where $B(\mu, L')$ defines a ball of radius L' around the point μ). Assuming that $\mu \leq L$, $B(\mu, L') \subset B(0, L + L')$ so that with $M = L + L'$, $N_{\mu, V}(B(0, M)) \geq 1 - \epsilon$ for all μ such that $\|\mu\| \leq L$. Choose $N \geq 1$ such that $M_n \geq M$ for all $n \geq N$. Let $n \geq N$ be given. Then:

$$\begin{aligned} P_0^n(\Phi_n(\mathbb{R}^d \setminus B(0, M_n)) > \epsilon) &\leq P_0^n(A_n) + P_0^n(\{\Phi_n(\mathbb{R}^d \setminus B(0, M_n)) > \epsilon\} \cap A_n^c) \\ &\leq \delta + P_0^n(\{N_{\Delta_n, V}(B(0, M_n))^c > \epsilon\} \cap A_n^c). \end{aligned}$$

Note that on the complement of A_n , $\|\Delta_n\| < L$, so:

$$N_{\Delta_n, V}(B(0, M_n))^c \leq 1 - N_{\Delta_n, V}(B(0, M)) \leq 1 - \inf_{\|\mu\| \leq L} N_{\mu, V}(B(0, M)) \leq \epsilon,$$

and we conclude that the last term on the *r.h.s.* of the previous display equals zero. \square

References

- [1] S. BERNSTEIN, *Theory of probability*, Moskow (1917).
- [2] R. BERK, *Limiting behaviour of posterior distributions when the model is incorrect*, Ann. Math. Statist. **37** (1966), 51–58. [MR0189176](#)
- [3] R. BERK, *Consistency of a posteriori*, Ann. Math. Statist. **41** (1970), 894–906. [MR0266356](#)
- [4] P. BICKEL and J. YAHAV, *Some contributions to the asymptotic theory of Bayes solutions*, Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete **11** (1969), 257–276. [MR0242298](#)
- [5] O. BUNKE and X. MILHAUD, *Asymptotic behavior of Bayes estimates under possibly incorrect models*, Ann. Statist. **26** (1998), 617–644. [MR1626075](#)
- [6] A. DAWID, *On the limiting normality of posterior distribution*, Proc. Canad. Phil. Soc. **B67** (1970), 625–633. [MR0258170](#)

- [7] S. GHOSAL, J. GHOSH and A. VAN DER VAART, *Convergence rates of posterior distributions*, Ann. Statist. **28** (2000), 500–531. [MR1790007](#)
- [8] P. HUBER, *The behavior of maximum likelihood estimates under nonstandard conditions*, Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability **1**, 221–233. University of California Press, Berkeley (1967). [MR0216620](#)
- [9] B. KLEIJN and A. VAN DER VAART, *Misspecification in Infinite-Dimensional Bayesian Statistics*. Ann. Statist. **34** (2006), 837–877. [MR2283395](#)
- [10] P. LAPLACE, *Théorie Analytique des Probabilités (3rd edition)*, Courcier, Paris (1820).
- [11] L. LE CAM, *On some asymptotic properties of maximum-likelihood estimates and related Bayes estimates*, University of California Publications in Statistics **1** (1953), 277–330. [MR0054913](#)
- [12] L. LE CAM and G. YANG, *Asymptotics in Statistics: some basic concepts*, Springer, New York (1990). [MR1066869](#)
- [13] E.L. LEHMANN. *Theory of Point Estimation*, John Wiley & Sons, New York. [MR0702834](#)
- [14] R. VON MISES, *Wahrscheinlichkeitsrechnung*, Springer Verlag, Berlin (1931).
- [15] D. POLLARD, *Convergence of Stochastic Processes*, Springer, New York (1984). [MR0762984](#)
- [16] L. SCHWARTZ, *On Bayes procedures*, Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete **4** (1965), 10–26. [MR0184378](#)
- [17] X. SHEN and L. WASSERMAN, *Rates of convergence of posterior distributions*, Ann. Statist. **29** (2001), 687–714. [MR1865337](#)
- [18] A. VAN DER VAART, *Asymptotic Statistics*, Cambridge University Press, Cambridge (1998). [MR1652247](#)
- [19] A. WALKER, *On the asymptotic behaviour of posterior distributions*, J. Roy. Statist. Soc. **B31** (1969), 80–88. [MR0269000](#)